Rapport détaillé sur l’analyse de la base de données des élections présidentielles en France en 2022

Ce rapport analyse les résultats de l'élection présidentielle française de 2022 à l'échelle régionale.
Les résultats seront expliqués en 3 parties : **I) Analyse descriptive simple** **II) AFC** **III) CAH**

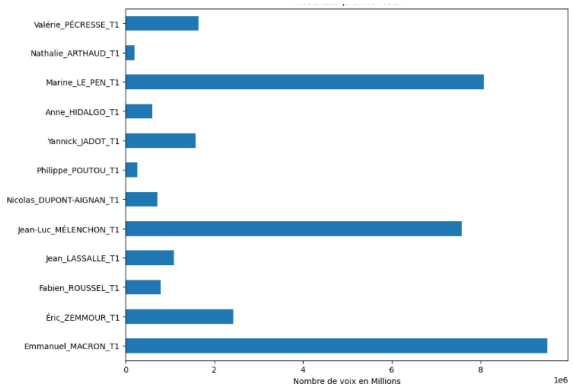
L'enjeu de la première partie est de proposer une étude descriptive complète du jeu de données afin, dans un premier temps, d'identifier les structures de vote du premier et second tour, puis de les mettre en relation afin de mettre en valeur de potentiels liens entre ces deux tours.

L'enjeu de la deuxième partie est d'approfondir l'analyse descriptive avec la méthode de l'AFC. Cette méthode permet de synthétiser l'information contenue dans nos données afin de visualiser graphiquement les grandes tendances (proximités et oppositions) entre les candidats et les profils régionaux.
Enfin, une CAH en 3ème partie permettra de regrouper les régions en classes homogènes ("clusters") en fonction des données du premier tour. Cela servira à bien représenter la structure du vote en mettant en évidence des groupes de territoires ayant les mêmes préférences, confirmant notamment la fracture entre l'Hexagone et les départements d'Outre-mer.

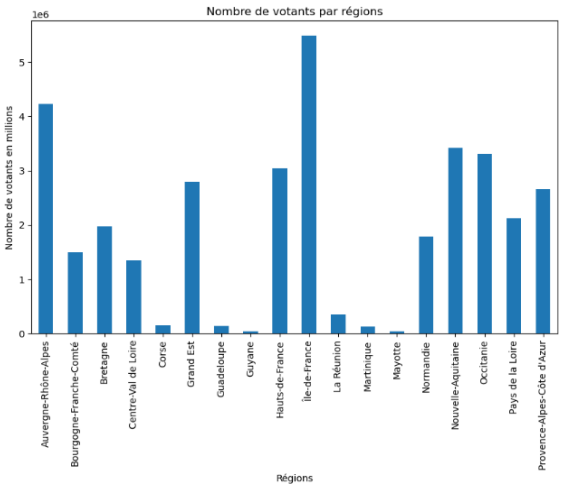
Partie 1 : Analyse Descriptive

Cette partie est assez libre. Nous avons décidé, premièrement, d'observer les votes globaux du premier tour (**GRAPHIQUE 1**), ce qui montre la domination du trio Macron, Le Pen et Mélenchon. Cependant, ces résultats nationaux sont indissociables du nombre d'habitants constituant les différentes régions. Le (**GRAPHIQUE 2)** montre que les résultats sont portés par quelques grandes régions (notamment l'Île-de-France), ce qui peut masquer des réalités locales très différentes.

GRAPHIQUE 1 :

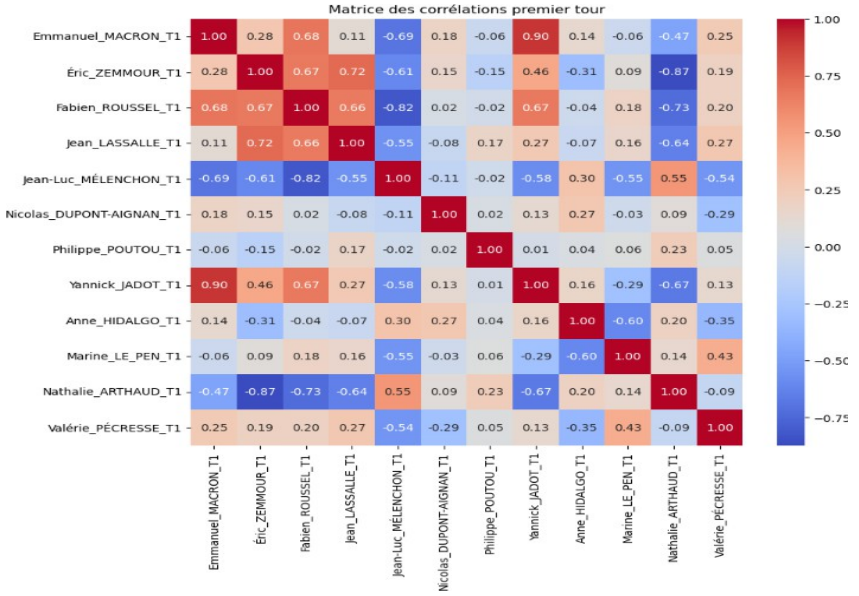


GRAPHIQUE 2:



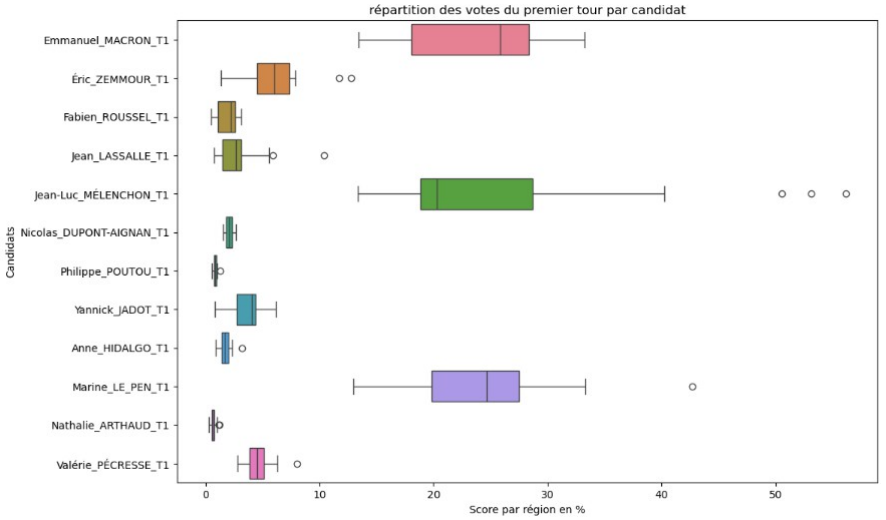
Après avoir normalisé les données par rapport au nombre d'habitants, nous remarquons que le **GRAPHIQUE 3** révèle une forte polarisation. Les corrélations négatives entre les 3 candidats de tête indiquent que leurs bases électorales occupent des espaces géographiques différents, il y a une séparation géographique nette entre les différents blocs de votants.

GRAPHIQUE 3 :



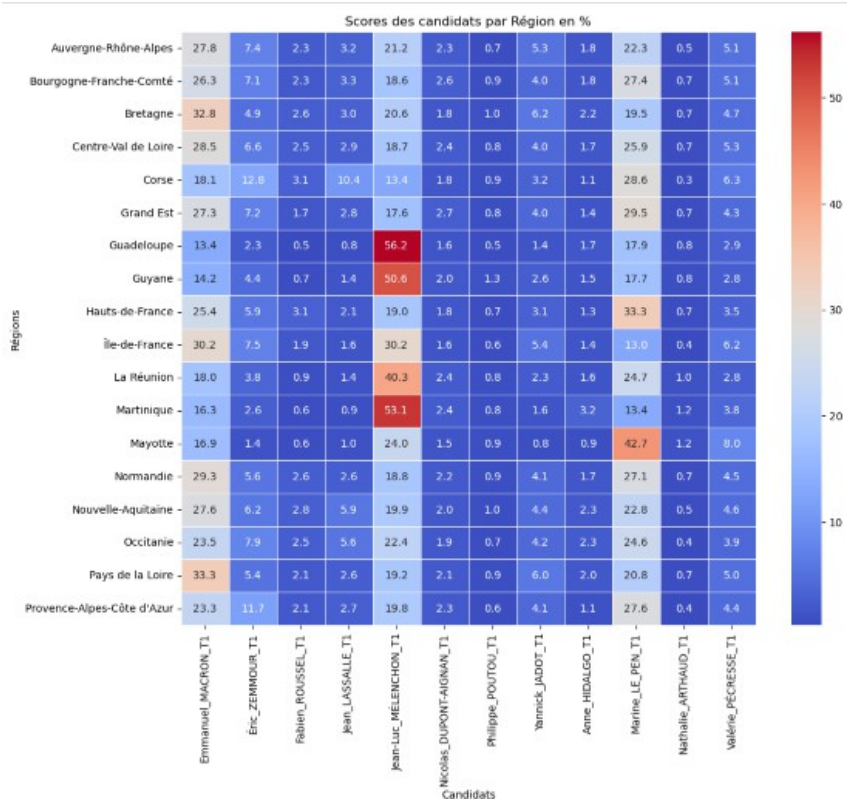
L'étude de la répartition par les boxplots (**GRAPHIQUE 4**) permet de voir que le soutien des candidats ne se répartit pas du tout de la même façon sur le territoire. Les favoris ont des boîtes très étirées, ce qui montre qu'ils possèdent des régions "fortes", leurs scores sont très variables d'une région à l'autre. Pour les autres candidats c'est l'inverse, sans réelle zone "forte", leurs résultats sont uniformes partout en France.

GRAPHIQUE 4 :



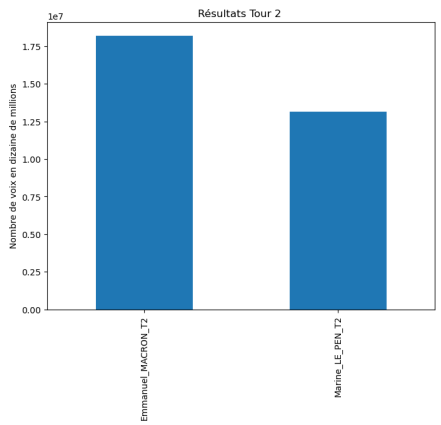
Ces zones "fortes" sont rendues explicites par le **GRAPHIQUE 5**, on y identifie clairement la domination de Mélenchon dans les zones urbaines et en outre-mer, tandis que l'extrême droite consolide ses positions dans le Nord (Hauts-de-France). Le candidat du centre possède une répartition plus homogène avec beaucoup de zones "assez-fortes" (Pays de la Loire, Bretagne).

GRAPHIQUE 5 :

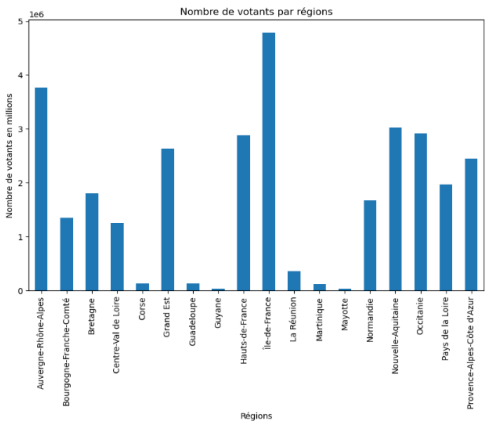


Deuxièmement, au second tour, le duel final confirme une victoire en volume pour Emmanuel Macron (**GRAPHIQUE 6**), mais l'électorat reste, là encore, très inégal selon les régions (**GRAPHIQUE 7**).

GRAPHIQUE 6 :

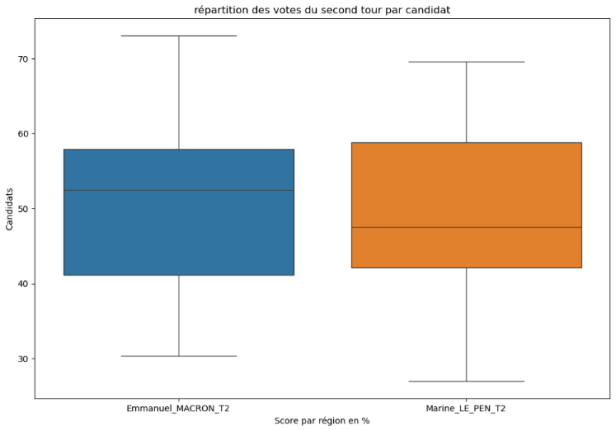


GRAPHIQUE 7 :

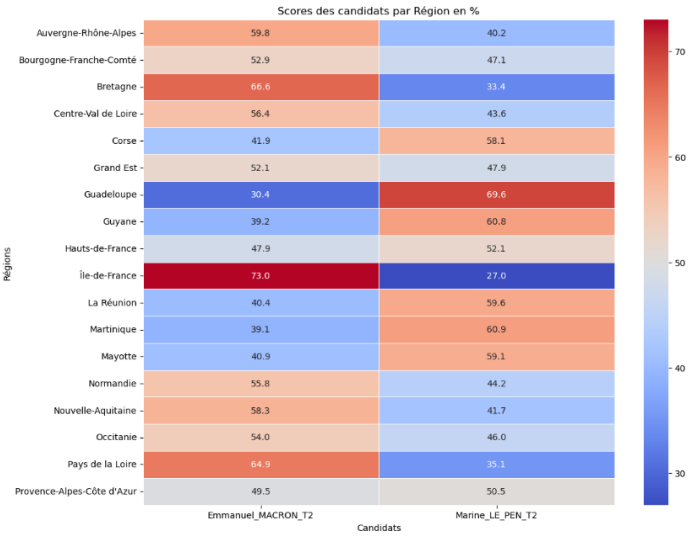


Comme précédemment, après avoir normalisé les données et afficher les boxplots, nous remarquons tout de suite l'amplitude des scores par région (**GRAPHIQUE 8**). L'écart entre les minimums et les maximums pour les deux candidats montre que la France est scindée en deux blocs aux visions opposées. Cet écart est confirmé par le **GRAPHIQUE 9**, les zones de rejet d'un candidat coïncident presque systématiquement avec les places fortes de son adversaire. Quelques régions conservent un certain équilibre, mais la tendance générale est celle de régions basculants très nettement d'un côté ou de l'autre.

GRAPHIQUE 8 :

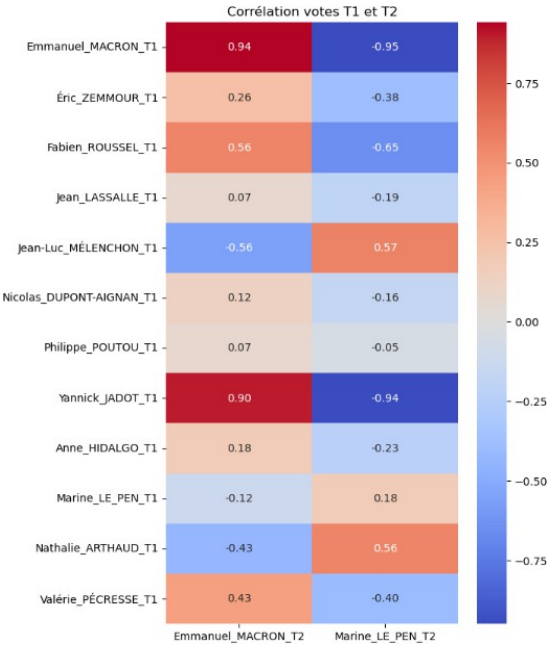


GRAPHIQUE 9 :



Enfin, le **GRAPHIQUE 10** est essentiel, il permet de comprendre comment les voix se sont déplacées entre les deux tours.

GRAPHIQUE 10 :



D'abord, on remarque une stabilité pour Emmanuel Macron, la corrélation entre son premier tour et son second tour (0.94) montre que son électorat est resté très solide sur ses bases géographiques habituelles. On voit aussi que l'électorat de Yannick Jadot (0,90) est presque identique à celui du président sortant au second tour.

Un résultat plus surprenant est que le vote Marine Le Pen au second tour est négativement lié à celui d'Éric Zemmour (-0,38), mais positivement à celui de Jean-Luc Mélenchon (0,57). On pourrait expliquer cela par un basculement, les régions qui avaient choisi Mélenchon au premier tour (comme l'Outre-mer, voir le **GRAPHIQUE 9**) ont massivement voté Le Pen au second par rejet du président sortant. À l'inverse, les zones "fortes" de Zemmour correspondent à des territoires où Emmanuel Macron a mieux résisté au second tour.

Le second tour montre ainsi une France coupée en deux entre les régions restées fidèles à Macron et les régions qui ont basculé de Mélenchon vers Le Pen par rejet du gouvernement.

En conclusion, cette première analyse montre que la géographie explique une grande partie du vote de 2022. On est passé d'un premier tour fragmenté en trois blocs à un second tour marqué par une rupture territoriale entre la continuité et la volonté de changement radical.

Partie 2 : Analyse Factorielle des Correspondances (AFC)

L'**Analyse Factorielle des Correspondances (AFC)** est une méthode d'analyse multivariée utilisée pour explorer les relations entre **des variables qualitatives** présentées sous forme d'un **tableau de contingence (ou tableau croisé)**.

Elle vise à :

- **Résumer l'information** contenue dans un tableau souvent très grand,
- **Réduire la dimension** tout en conservant les structures d'association entre les modalités,

- **Visualiser graphiquement** les liens ou oppositions entre les lignes (profils des individus ou groupes, ici les régions) et les colonnes (catégories ou modalité d'une autre variable, ici les candidats).

Autrement dit, l'AFC permet de représenter **dans un même plan** les modalités des deux variables, afin d'identifier **les proximités** et **les oppositions** entre elles.

Pourquoi c'est intéressant dans notre cas ?

Il est pertinent de réaliser une AFC car nous avons en données un tableau de contingence avec 2 variables qualitatives, le tout avec de nombreuses modalités (et donc beaucoup de dimensions), et en addition les données du tableau sont des effectifs (chaque case représentant le nombre de votes pour une région et un candidat donné).

Tout ceci est parfait pour une AFC qui va permettre de mieux synthétiser et visualiser l'information et les liens entre les différents profils lignes et colonnes en réduisant les dimensions de notre tableau sous forme d'axes principaux et donc de plans bidimensionnels.

De plus, on a vu dans la 1ère partie d'analyse descriptive qu'on voit apparaître des liens de corrélations entre les différentes modalités (notamment une séparation géographique nette entre les différents blocs de votants).

Il est donc pertinent de réaliser l'AFC pour mieux identifier et analyser les correspondances qu'on a vu dans la 1ère partie entre certaines régions et candidats et de mettre en évidence des sous-groupes de régions ou de candidats avec les mêmes profils.

Vérification de l'intérêt de la méthode sur nos données via le test Khi-deux :

On vérifie tout de même la pertinence de l'AFC avec le test du Khi-Deux, qui est une étape préalable fondamentale pour le calcul des axes et la vérification de l'utilité de l'application de l'AFC sur nos données.

Il vérifie si les variables qualitatives (lignes et colonnes d'un tableau de contingence) sont indépendantes ou associées.

Si les variables sont indépendantes (hypothèse H0) → aucune structure particulière à explorer, l'AFC n'apporterait pas d'information pertinentes.

Si les variables sont dépendantes (hypothèse H1) → il existe des liens entre les modalités, et l'AFC devient pertinente pour les visualiser et les interpréter dans un espace réduit.

On réalise le test Khi deux et l'AFC sur les données du 1er tour seulement car les données du 2nd tour ne contiennent que 2 candidats => peu de dimensions, pas très intéressant de le faire (on garde les résultats de notre analyse descriptives sur le 2nd tour)

Résultats : Statistique chi2 très élevée (donc grosse différence entre effectifs réelles et effectifs attendues si 0 indépendance entre les variables) et p-value extrêmement faible ou nulle alors que seuil à ne pas dépasser = 5% =) Dépendance entre modalités reconnues, l'AFC est utile.

Application AFC :

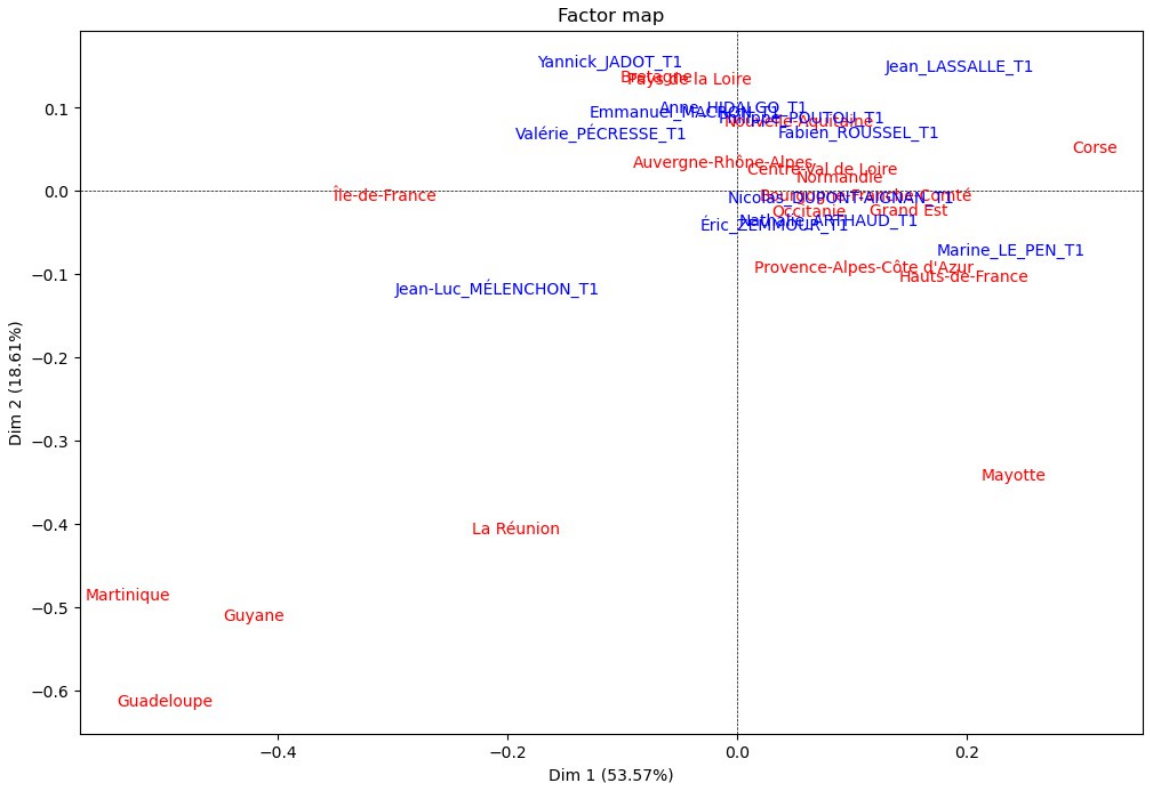
On trouve les valeurs propres ainsi que le % de variance des axes principaux qui sont les combinaisons linéaires des colonnes ou des lignes (ici des régions et candidats) qui sont choisies pour maximiser l'inertie/ la variance expliquée des données.

Les 2 premiers axes expliquent + de 70% de l'inertie/variance des données, ce qui est suffisant pour se concentrer sur leur étude.

On obtient en appliquant la méthode :

1. La contribution de chaque modalité pour chaque axe (la part de variance de l'axe expliquée par la modalité, ici d'inertie car AFC et pas ACP).
2. Le cosinus carré de chaque modalité pour chaque axe (si la modalité est bien représentée dans l'axe).

A partir de ça et du nuage de point suivant, on va pouvoir interpréter les liens entre les régions et les candidats :



Interprétation nuages de points :

L'axe 1, qui explique la plus grande part de l'inertie (53.57%), oppose principalement deux profils régionaux très différents.

D'un côté, on observe la candidature de Jean-Luc Mélenchon associée à la région Île-de-France, qui présente des contributions et des cos² élevés, indiquant qu'ils structurent fortement cet axe et qu'ils sont bien représentés.

De l'autre côté de l'axe apparaît la candidature de Marine Le Pen, également très contributive mais située en opposition sur l'axe.

=> Côté droit : Le Pen associée à des régions comme les Hauts de France et la côte d'Azur.

=>Côté gauche : Mélenchon associé à des régions comme l'Ile de France et globalement la plupart des régions ultramarines (Réunion, Guyanne…).

L'axe 2, qui explique une part moindre mais encore significative de l'inertie (18.61%), révèle une structuration complémentaire des données.

Il est principalement porté par les candidatures d'Emmanuel Macron et de Yannick Jadot, qui présentent une bonne qualité de représentation ainsi qu'une forte contribution sur cet axe.

Plusieurs régions d'outre-mer, telles que La Réunion, la Guadeloupe et la Martinique, ainsi que certaines régions de l'Ouest de la France comme la Bretagne et les Pays de la Loire, contribuent également fortement à cet axe.

=> En haut : Macron et Jadot associés fortement à des régions comme la Bretagne et le Pays de la Loire.

Conclusion :

On retrouve quasiment les mêmes conclusions que celles de notre analyse descriptive initiale :

On a nos 3 favoris (Mélenchon, Lepen et Macron) qui sont surreprésentés dans certaines régions et ont des zones fortes.

Tandis que Mélenchon va avoir beaucoup de succès dans les territoires ultramarins ainsi qu'en Ile de France, Le Pen va avoir plus de succès et de voix dans d'autres régions, dont certaines plus au Nord (Haut de France, Grand Est, Corse). Macron, lui, s'en sort bien dans pas mal de régions et à des zones fortes dans des pays comme la Bretagne et le pays de la Loire.

Partie 3 : Classification Ascendante Hiérarchique (CAH)

La CAH est une méthode d'analyses multivariées qui permet de regrouper des individus (régions) en groupes homogènes appelés « clusters ».
Le principe ascendant est visualisé à travers le dendrogramme qui permet de regrouper les individus qui se ressemblent le plus (les plus proches). Et ainsi de suite jusqu'à ce qu'il ne reste plus qu'un groupe.

On utilisera que les données du premier tour. On étudiera donc les douze candidats du premier tour des présidentielles françaises de 2022.
Le but de l'analyse par CAH est d'analyser la structure de vote par régions. On cherche donc à croiser des régions (lignes) avec des candidats (colonnes). Autrement dit savoir quelles régions ont voté majoritairement pour quels candidats.

Le choix d'une CAH sur le premier tour s'explique simplement par le fait qu'au second tour, il n'y a que 2 candidats. L'analyse n'a pas beaucoup d'intérêt et donne lieu à un vote par "élimination", "barrage". Et agrège des bulletins dont les électeurs ne sont peut-être pas d'accord.

La CAH est effectuée sur les données normalisées (en % ici). Cela rend l'interprétation plus lisible, on peut dire que tel cluster est représenté à x % par tel candidats. Ce qui est plus compréhensible que les axes des AFC.De même l'AFC ne garde que 2 ou 3 axes, ce avec le plus de variance. Elle ne tient pas compte des candidats forts dans des régions spécifiques. La normalisation permet donc de conserver l'information à contrario de l'AFC.

L'analyse du dendrogramme montre que la structure du vote est très proche pour le cluster vert (Martinique, Guyane, Guadeloupe, La Réunion). Ce groupe est également le plus isolé sur l'arbre. Ce qui traduit une fracture géographique du vote entre l' Hexagone et les DROM. Ce cluster est lui-même très éloigné dans leurs suffrages de celui tout à gauche du Grand-Est à la Provence-Alpes-Côte d'Azur (PACA).

On remarquera que contrairement aux autres DROM, Mayotte est tout seul. C'est un outlier ou le vote pour Le Pen est un des plus importants.

Pour la découpe en cluster, on utilise la méthode du coude, on trace l'inertie reliant les clusters. Pour ne pas perdre le cas de Mayotte on choisit 4 clusters et non 3. Chaque cluster représente des régions avec un vote similaire. L'interprétation peut se faire soit à travers le nuage de points soit à travers le

tableau des moyennes des voix obtenues par chaque candidat découpé par **4 clusters**.

Cluster 0 : Martinique, Guyane, Guadeloupe, La Réunion.
Ces régions votent massivement pour Mélenchon, il confirme les hypothèses qu'on a pu établir avec le dendrogramme. Ils ont voté en moyenne pour Mélenchon avec près de 74 700 voix.
On peut supposer ce vote par les inégalités sociales entre ces territoires et la métropole et peut-être une volonté d'un état plus interventionniste.
Mélenchon est suivi par Le Pen avec en moyenne 33 200 voix et de Macron avec 26 455 voix.

Cluster 1 : Auvergne-Rhône-Alpes, Bretagne, Île-de-France, Nouvelle-Aquitaine, Pays de la Loire.
Ces régions votent massivement pour Macron. Il y récolte en moyenne 1 026 000 voix.
C'est principalement dans ces régions que se situent les métropoles, les services publics et privés.
Logiquement on y retrouve les classes les plus aisées dont les cadres et autres CSP+ dont le vote est pour Macron d'après de nombreux sondages.
Macron est suivi par Mélenchon avec en moyenne 810 000 voix et Le Pen avec 652 340 voix.

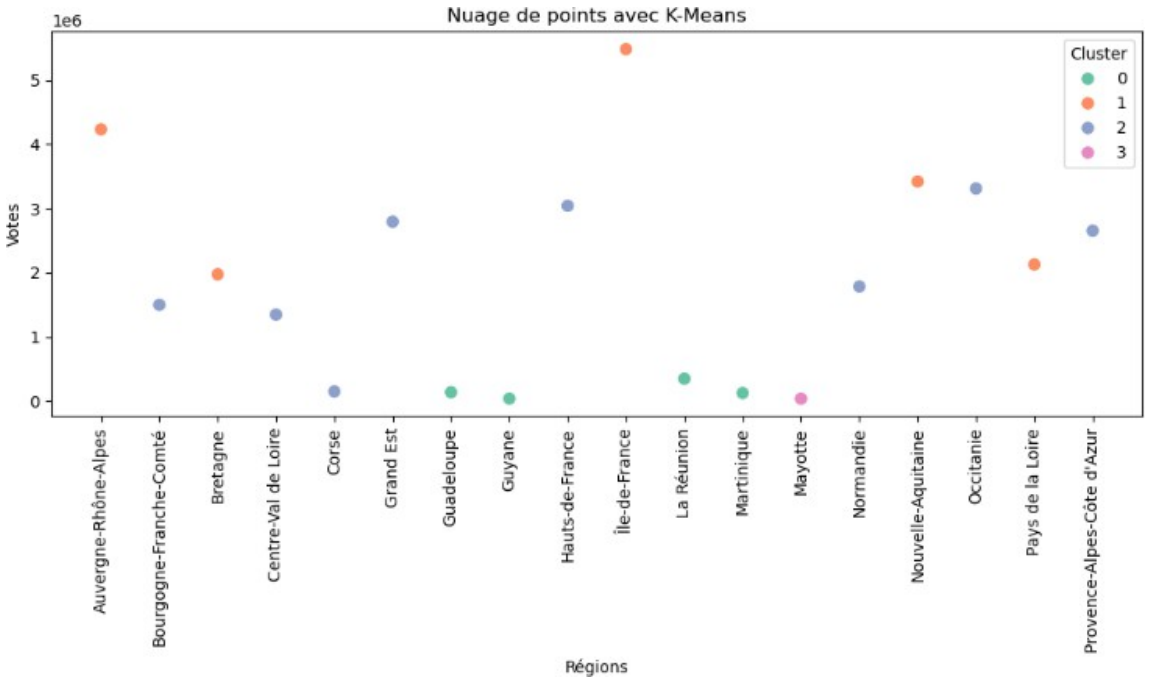
Cluster 2 : Bourgogne-Franche-Comté, Centre-Val de Loire, Corse, Grand-Est, Hauts-de-France, Normandie, PACA.

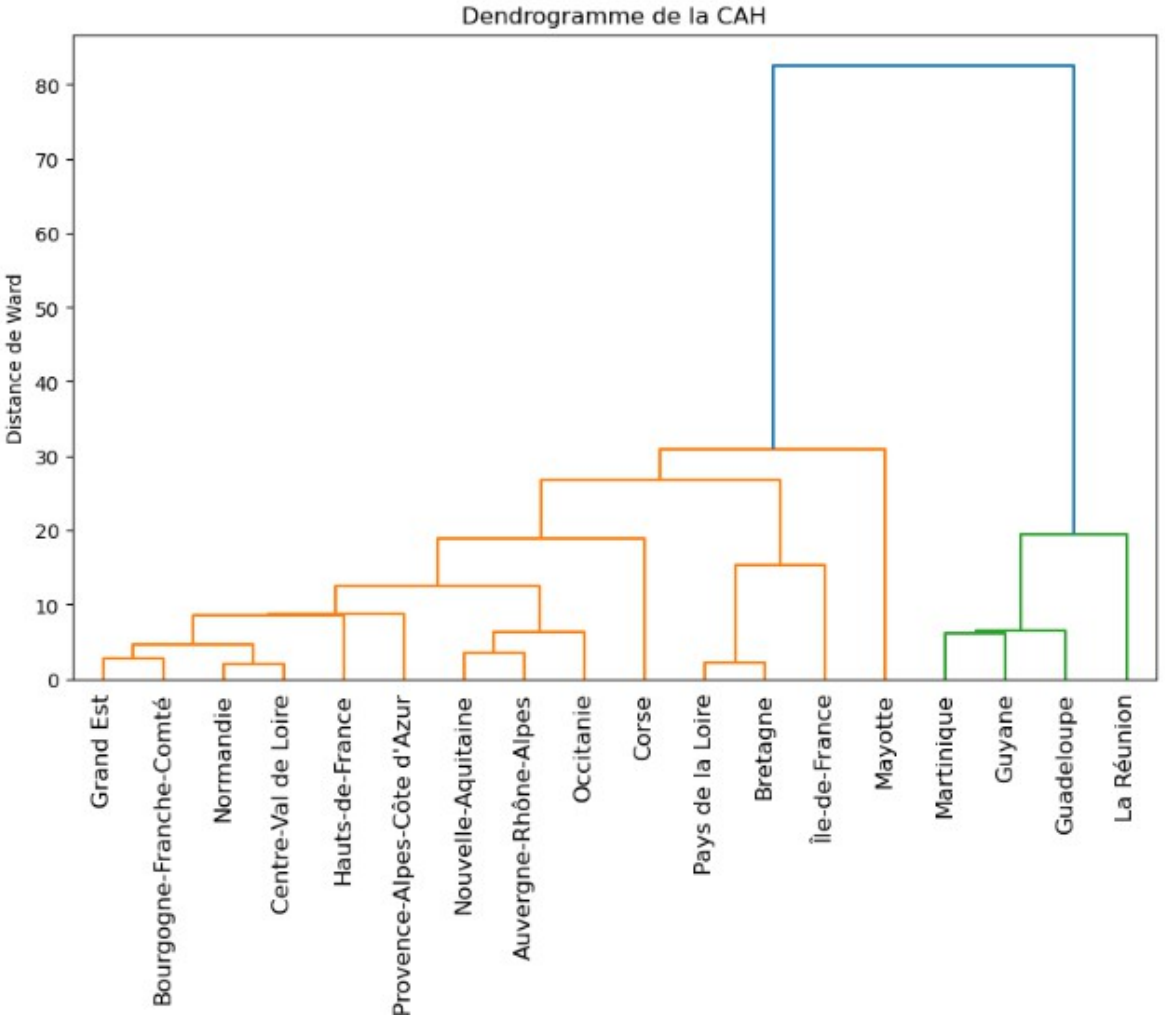
Ce cluster est mené par Le Pen avec une moyenne de 583 815 voix.
Ce cluster représente une France des « périphéries ». On y compte de grandes régions agricoles et industrielles (ou anciennes) telles que le Grand-Est ou les Hauts de France.
Le Pen est suivi par Macron avec en moyenne 532 252 voix puis par Mélenchon avec 402 562 voix.

Cluster 3 : Mayotte.

Ce cluster est dominé par un vote Le Pen avec 14 958 des voix exprimées en 2022 (pas une moyenne).
Ce cluster atypique souligne le cas particulier de ce DROM confronté à une situation migratoire particulière avec les îles voisines de Comores depuis plusieurs années.
Le Pen est suivi par Mélenchon avec 8 398 voix puis par Macron avec 5 936 voix.

Il est intéressant de noter que Mélenchon est le second candidat après Le Pen pour ce cluster.
En effet, on oppose là une candidate qui est opposé à l'immigration clandestine face à un autre candidat qui souhaite régulariser les sans-papiers.


Nuage de points avec K-Means
1e6
Votes
Cluster
0
1
2
3
Auvergne-Rhône-Alpes
Bourgogne-Franche-Comté
Bretagne
Centre-Val de Loire
Corse
Grand Est
Guadeloupe
Guyane
Hauts-de-France
Île-de-France
La Réunion
Martinique
Mayotte
Normandie
Nouvelle-Aquitaine
Occitanie
Pays de la Loire
Provence-Alpes-Côte d'Azur
Régions


Dendrogramme de la CAH
Distance de Ward
Grand Est
Bourgogne-Franche-Comté
Normandie
Centre-Val de Loire
Hauts-de-France
Provence-Alpes-Côte d'Azur
Nouvelle-Aquitaine
Auvergne-Rhône-Alpes
Occitanie
Corse
Pays de la Loire
Bretagne
Île-de-France
Mayotte
Martinique
Guyane
Guadeloupe
La Réunion

Synthèse des 3 parties :

Ce travail met en évidence que le vote de 2022 est avant tout structuré par une fracture territoriale majeure. L'analyse descriptive a d'abord montré la nécessité absolue de normaliser les données pour contourner le poids démographique des grandes régions, révélant ainsi une forte stabilité du vote Macron entre les deux tours qui contraste avec l'éclatement de l'électorat Mélenchon. L'AFC valide l'espace électoral autour de deux axes majeurs : le premier oppose nettement le vote Mélenchon (Île-de-France, Outre-mer) au vote Le Pen (Hauts-de-France, PACA), tandis que le second isole la spécificité du vote Macron ancré dans l'Ouest. Enfin, la classification (CAH) objective ces observations en découpant la France en quatre zones distinctes : un cluster "DROM" (Martinique, Guadeloupe, Guyane, La Réunion) votant massivement pour Jean-Luc Mélenchon par interventionnisme social , un cluster "Métropolitain et de l'Ouest" aisé et légitimiste (CSP+) acquis à Emmanuel Macron , un cluster des "Périphéries" (Grand-Est, Hauts-de-France, PACA) dominé par le vote Le Pen , et enfin l'exception de Mayotte qui forme un groupe à part en plébiscitant Marine Le Pen pour des raisons migratoires spécifiques. Cette typologie confirme que la géographie et les spécificités locales restent les déterminants principaux du comportement électoral.